基于句子跨度的哈萨克语句法分析研究 *

柴 伟[1,2,3]，古丽拉·阿东别克[1,2,3]

(1. 新疆大学 信息科学与工程学院，乌鲁木齐 830046; 2. 新疆多语种信息技术实验室，乌鲁木齐 830046; 3. 国家语言资源监测与研究少数民族语言中心哈萨克和柯尔克孜语文基地，乌鲁木齐 830046)

摘 要：由于目前哈萨克语句法分析准确率较低并缺乏基于神经网络的哈萨克语句法分析的相关研究，针对哈萨克语短语结构的句法分析，使用基于移进-归约的方法，但采用在栈中存储句子跨度而不是部分树结构，从而在进行句法树解析时不需要对句法树进行二叉化。该研究在句子特征提取时使用双向 LSTM 对句子跨度特征进行提取，得到句子跨度在整个句子上下文中信息，再使用多层感知机对句法分析模型进行训练，最后在解码时使用动态规划选取最优句法分析结果；最终使得哈萨克语短语句法分析准确率达到了 76.92%。研究成果对哈萨克语句法分析准确率有了进一步的提高并为后续的哈萨克语机器翻译及语义分析奠定良好的基础。

关键词：双向 LSTM；句子跨度；动态规划

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2018.08.0632

Research of kazakh parsing based on span

Chai Wei[1,2,3], Gulila Altenbek[1,2,3]

(1. *College of Information Science & Engineering, Xinjiang University, Urumqi* 830046, *China*; 2. *Xinjiang Laboratory of Multi-language Information Technology, Urumqi* 830046, *China*; 3. *The Base of Kazakh & Kirghiz Language of National Language Resource Monitoring & Research Center on Minority Language, Urumqi* 830046, *China*)

Abstract: Due to the low accuracy of Kazakh parsing and the lack of correlation research based on neural network Kazakh parsing. This paper focused on the parsing of Kazakh phrase structure, based on the shift-reduce method, but by the stack elements were sentence spans rather than partial tree, then it didn't need to carry out the binary tree in parsing. It also used the bi-directional LSTM to extract the features of sentence span, and obtained the sentence span in the whole sentence context, using the multilayer perceptron to train the parsing model. In the end, the Kazakh parsing accuracy has been achieved 76.92%. The research results have improved the accuracy of Kazakh parsing and build a good foundation for Kazakh machine translation and semantic analysis.

Key words: Bi-LSTM; span; dynamic oracle

0 引言

句法分析是根据给定的语法体系，自动推导出句子的语法结构，分析句子所包含的语法单元和这些语法单元之间的关系，将句子转换为一棵结构化的语法树[1]。句法分析可为更深层次的机器翻译、语义角色标注等提供良好的基础，是自然语言处理中起承上启下的重要作用的一个研究环节。

在句法分析研究中，基于规则的方法进行句法分析研究时，规则由语言学家进行提炼，例如 Collins、Hallet 等人[2,3]使用上下文无关文法产生的规则或根据产生的规则对句子进行打分，生成短语句法树。在基于统计的句法分析研究中，例如：Charniak[4]提出了单纯的 PCFG 的句法分析方法，并在此基础上又提出了基于词汇化的 PCFG 句法分析方法。目前，使用神经网络的方法对句法分析准确率有较大提高，语法规则在句法分析中的重要性降低。例如 Dyer、Stern[5,6]等人使用编码—解码模型，编码主要是将句子读入并表示为向量，解码是使用这些向量生成标记好的句法树。

哈萨克语的自然语言处理目前已经进行了短语结构的句法分析研究，文献[7]利用 PCFG 方法，并将 Viterbi 算法加入解码模块，实现了有自学习能力的哈萨克语短语句法分析器。文献[8]将概率上下文无关文法模型和 Chart 算法结合，实现了基于 PChart 算法的哈萨克语短语句法分析器；文献[9]中使用 PCFG 与感知机结合进行重排序，实现了由粗到精的哈萨克语短语句法分析等。

虽然哈萨克语句法分析已在基于规则和统计的方法上进行了一定的研究，但随着神经网络技术的发展，对于句子序列预测的模型越来越多，准确率也越来越高，面临此类技术应用到哈萨克语句法分析中问题。本文在基于移进—归约系统基础上，将栈中存储的部分树结构改为存储句子跨度，使用双向 LSTM 获取句子跨度的上下文特征。双向 LSTM 获取的是句子级别的信息，得到句子跨度在整个句子上下文中信息。同时也将动态规划方法应用在哈萨克语句法分析中，使得在句法分析解码时不需要重排序就可取得句法分析的最好的结果。句子跨度相较于部分树结构的优势主要是：句子跨度表示句子的一个片段对应的子树，核心词必须位于段首或段尾，因此减少了解码时的搜索空间，使搜索空间复杂度降低。

本文的主要贡献有：a）将神经网络应用在哈萨克语句法分析中，并取得了良好的效果；b）进行哈萨克语短语句法分析时将句子跨度作为移进—归约方法中栈中存储的基本单

收稿日期：2018-08-29；**修回日期：**2018-11-01 **基金项目：**国家自然科学基金资助项目（61363062）

作者简介：柴伟（1994-），女，新疆昌吉人，硕士，主要研究方向为自然语言处理（chai1012@foxmail.com）；古丽拉·阿东别克（1962-），女，新疆乌鲁木齐人，教授，博导，主要研究方向为自然语言信息处理.

位；c）使用双向 LSTM 神经网络获取句子上下文信息；d）使用动态规划使得句法分析解码时不需要重排序就可得到最好的句法分析结果。

1　哈萨克语句法分析系统

哈萨克语属阿尔泰语系中的突厥语族，是黏着语，根据哈萨克语自身的特点目前划分了哈萨克语二级词组标注 5 个基本短语和 10 个一级词性标注[8]。

在文献[10,11]中提出了一种句法分析的新方法，主要依赖于从句子表面跨度直接提取信息，而不是通过丰富的语法结构来获得信息。短语结构的句法树可以被看做是一个句子中有标记的跨度的集合。以此为指导原则，本文提出了一个模型，它由两个部分组成，一个部分给句子跨度标签分配打分，另一个部分直接对句子跨度是否存在进行打分。前者用于确定输出的标记，后者提供输出的句子跨度结构。这两部分核心问题是句子跨度的表示问题。由于一个句子跨度的正确标记在很大程度上取决于它出现的上下文，所以将这一特点与神经网络联系起来，神经网络已经被证明能够获得适合于各种自然语言的上下文信息。

在整个句法分析的过程中，是以基于移进-归约的方法作为基本模型，主要涉及到两种数据结构：输入队列和分析栈。其中 σ 表示分析栈，其中在栈中存储的是自下而上进行句法分析的句子短语对应的句子跨度；z 为句法分析动作数，记录移进-归约句法分析的步骤数；i，j 为栈中待分析元素；X 为一元链或者非终端节点，只有一次标记动作的选择，所以对于 n 个单词组成的句子只需要（4n-2）个步骤就可以完成句法分析；t 表示句子跨度的集合，在句法树解析完成时，存放整个句法树。句法解析过程中主要有四种动作：

a）移进（shift）。将队列中的<词，词性>压入栈中；

$$\frac{z,\sigma|j,t}{z+1,\sigma|j|j+1,t},\quad j<n,\ \ even\ z$$

b）归约(comb)。结合栈顶的两个句子跨度。

$$\frac{z,\sigma|i|k|j,t}{z+1,\sigma|i|j,t}\quad,\qquad even\ z$$

c）标记栈顶(label-X)。

$$\frac{z,\sigma|i|j,t}{z+1,\sigma|i|j,t\cup\{{}_iX_j\}},\qquad odd\ z$$

d）不进行标记(nolabel)。

$$\frac{z,\sigma|i|j,t}{z+1,\sigma|i|j,t},\ z<(4n-1)\quad odd\ z$$

以上四种动作的形式化表示中，横线上方表示当前“焦点词”状态，横线下方表示执行相应分析动作之后的结果。前两种称为结构动作，后两种称为短语标记动作。短语标记动作主要是将栈中一个或更多的非终端节点的句子跨度转换成部分树结构。

以上的方法与文献[12]所提出的奇偶句法分析方法相似，但由于在栈中存储的不是部分树结构，在训练和句法树解析时都不需要进行二叉化。例如，给出一个句子：

ءبىز ۇرىمجىگە ۇشاقپەن ۇشىپ كەلدىك .

图 1 为此句子的标准句法树，定义 $i\,\Delta\,j$ 表示句子跨度(i，j)，如表 1 所示，步骤 1-2 进行移进动作，由于不是一个短语，就不进行标记(nolabel)；步骤 5-6，栈顶的两个句子跨度可以组成短语，那么就进行归约动作，并进行短语标记（label-VP），其他步骤类似。以上只是使用静态解码方法生成“短堆栈”式的句法树，意味着移进和规约动作都可以产生时，总是归约优先。

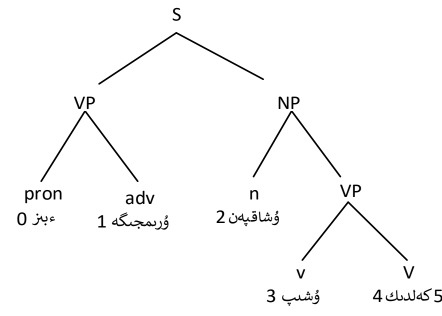


图 1　哈萨克语标准句法树

Fig. 1　Kazakh gold parse tree

表 1　静态解码动作

Table 1　Static decode actions

步数	结构动作	标记动作	栈	括号
1-2	shift(ءبىز/pron)	nolabel	$_0\Delta_1$	
3-4	shift(ۇرىمجىگە/adv)	nolabel	$_0\Delta_1\Delta_2$	
5-6	comb	label-VP	$_0\Delta_2$	$_0VP_2$
7-8	shift(ۇشاقپەن/n)	nolabel	$_0\Delta_2\Delta_3$	
9-10	shift(ۇشىپ/v)	nolabel	$_0\Delta_2\Delta_3\Delta_4$	
11-12	shift(كەلدىك/v)	nolabel	$_0\Delta_2\Delta_3\Delta_4\Delta_5$	
13-14	comb	label-VP	$_0\Delta_2\Delta_3\Delta_5$	$_3VP_5$
15-16	comb	label-NP	$_0\Delta_2\Delta_5$	$_2NP_5$
17-18	comb	label-S	$_0\Delta_5$	$_0S_5$

句法分析的输入是词和词性的向量，词性是事先标注好的作为句法分析的输入。词向量的初始化可以是随机初始化或是使用大规模语料库事先训练好。在本文中使用随机初始化词向量。采用的神经网络结构如图 2 所示：在使用 2 层的双向 LSTM 获得句子跨度特征后，使用多层感知机进行训练，将每个单词通过 d 维向量表示 $e_i^w\in\boldsymbol{R}^d$ ；同样，将词性也用 d 维向量表示 $e_i^l\in\boldsymbol{R}^d$ 。将 $x^w=[e_{w_1}^w e_{w_2}^w\cdots e_{w_{n_w}}^w]$ 以及 $x^l=[e_{w_1}^l e_{w_2}^l\cdots e_{w_{n_w}}^l]$ 作为输入，通过 ReLU 线性激活函数将输入单词和词性数据分别映射到隐藏层，其中 $W_1^s\in\boldsymbol{R}^{dh\times(d\cdot n_w)}$ ，$W_1^l\in\boldsymbol{R}^{dh\times(d\cdot n_l)}$；$b_1^s$，$b_1^l\in\boldsymbol{R}^{dh}$ 是偏置项。

$$h_s=\max\{0,W_1^s x^w+b_1^s\}\tag{1}$$

$$h_l=\max\{0,W_1^l x^l+b_1^l\}\tag{2}$$

在隐藏层最上面添加一个 softmax 层计算多类概率，得出最后的标记结果的最大概率，其中 $W_2^s\in\boldsymbol{R}^{2\times dh}$ ，$W_2^l\in\boldsymbol{R}^{|L|\times dh}$ ，$|L|$为句子短语标记的数量。

$$p_s=\mathrm{softmax}(W_2^s h_s)\tag{3}$$

$$p_l=\mathrm{softmax}(W_2^l h_l)\tag{4}$$

2　句子跨度特征提取

在文献[13]中提出了使用一层的双向 LSTM 可以表示词在句子在中上下文中的位置信息；2 层的双向 LSTM 可以更进一步提高句法分析性能,并将两层位置特征表示的输出，成为全连接层的输入。本文使用 2 层双向 LSTM 来获得一个输入 i 在前向 f_i 和后向 b_i 编码后的上下文信息，句子跨度(i，j)

的表示是通过计算向量 $f_j - f_i$ 和向量 $b_i - b_j$ 的不同，并与移进-归约系统结合进行句法分析。如图 3 所示，句子跨度 3 ۋشىپ 4 كەلدىك 5 是通过计算向量 $(f_5 - f_3)$ 和向量 $(b_3 - b_5)$ 在这个模型中，句子输入一次，就可以将相同的循环输出用来计算句子跨度特征贯穿于整个句法分析的过程。

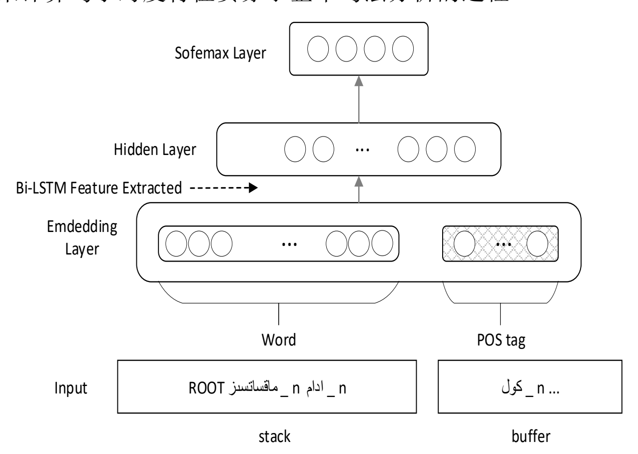


图 2　哈萨克语句法分析神经网络结构图

Fig. 2　Neural network architecture of Kazakh parsing system

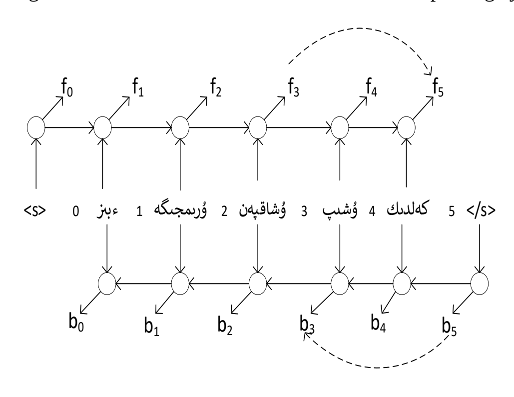


图 3　LSTM 获取句子跨度特征

Fig. 3　Word spans are model by differences in LSTM

当进行句法分析考虑到结构动作时，需要考虑是否将栈顶的两个句子跨度进行合并；在考虑标记动作时栈顶的句子跨度是动作标记的一个重要的候选。因此，使用 4 个句子跨度特征决定结构动作，3 个句子跨度特征决定短语标记动作，这样就可以精确划分句子，在这两种情况下同时也要考虑剩余自己的前缀和后缀。特征模板如表 2 所示。

表 2　句子跨度特征模板

Table 2　Span feature templete

动作	栈	句子跨度特征
结构动作	$\sigma \mid i \mid k \mid j$	0□$i\Delta k\Delta j$□n
标记动作	$\sigma \mid i \mid j$	0□$i\Delta j$□n

3　动态规划

在文献[14~16]中已经使用并证明了动态规划在标记跨度上的最优性。动态规划对于标记的决策主要有两种：

a)如果分配给句子跨度的标记是标准句法树的一部分，那么就是正确的句子跨度的标记。

b)如果是其他情况那么标记就为空。

动态规划在结构方面主要对句法树的分支点进行决策，如果句子跨度(i，j)在标准句法树中出现，那么 b(i，j)表示当前句子跨度的内部边界点，例如句子跨度(1,8)有子跨度(1,3)(3,5)(5,8)，那就有两个内部边界点 b(1,8)={3,5}，那么动态规划的作用就是将句子跨度与标准句法树一致的进行精确输出。如果句子跨度(i，j)与标准句法树不对应，就要识别与标准句法树相似的最小的句子跨度。

$$(i^*, j^*) = \min\{(i', j') \in T : i' < i < j < j'\} \tag{5}$$

其中最小值取决于由跨度长度引起的偏序。然后动态规划的输出就是这个包围跨度的内部边界点的集合，它也位于原始的内部边界点，$\{k \in b(i^*, j^*) : i < k < j\}$　。

4　实验

4.1　实验数据及评价指标

本文实验数据来自是《新疆日报》(哈萨克文版)，所使用的语料都是经过人工标注和校对，并将所有语料按照 8：1：1 的比例分配为训练集、验证集以及测试集。

实验原始数据格式为

ءبىز ۋرىمجىگە ۋشاقپەن ۋشىپ كەلدىك .

实验测试结果数据格式为

(TOP (VP (pron ءبىز) (adv ۋرىمجىگە)) (NP (n ۋشاقپەن) (VP (v ۋشىپ) (v كەلدىك)) (. .)))

在本实验中，按照 PARSEVAL 的句法评测标准主要计算标记准确率（precision），标记召回率（recall），*F*-值（F_1 -measure）来衡量句法分析结果的好坏。评价函数的定义如下所示：

准确率

$$LP = \frac{\text{分析得到的正确短语个数}}{\text{分析得到的短语总数}} \times 100\% \tag{6}$$

召回率

$$LR = \frac{\text{分析得到的正确短语个数}}{\text{标准树库中的短语个数}} \times 100\% \tag{7}$$

F 值

$$F_1 = \left(\frac{2 \times LP \times LR}{LP + LR}\right) \times 100\% \tag{8}$$

4.2　实验环境

本实验使用 python 语言进行程序设计，其中，使用的 LSTM 神经网络模型来源于 DyNet 神经网络工具包[17]。DyNet 是由 Google 以及卡内基梅隆等多所大学共同开发，是一种为自然语言处理提供动态神经网络的工具包，包含 RNN，CNN 以及 LSTM 等常用神经网络模型。

实验的硬件环境为 32 GB RAM，使用 GPU 为 NVIDIA GeForce GTX 1080Ti，操作系统为 64 位 Ubuntu16.04。

4.3　实验结果及分析

本文提出的哈萨克语句法分析模型中，对句法分析效果影响较大的因素为词向量特征以及句子长度。本文将设计以下实验进行测试。

4.3.1 LSTM 提取特征对句法分析准确率影响

使用 Bi-LSTM 对句子特征进行提取时，其隐层节点数量会影响到句子特征提取的质量。为了考察 Bi-LSTM 隐层节点数量对哈萨克语句法分析性能影响，通过改变 Bi-LSTM 隐层节点数量进行实验，隐层数量设置为 50、70、100、120、150、200，得到结果如图 4 所示。

从实验结果可以得出以下结论：a）隐层数量的大小会影响对特征的提取，进一步会影响句法分析的效果；b）隐层数量大小增加到 120 后，继续增加隐层数量对于句法分析效果的提升并不明显，所以在后续实验中设置隐层大小为 120。

4.3.2 句子长度对句法分析准确率影响

在句法分析中句子长度往往也会对句法分析效果产生影响，将实验语料中句子长度按照词的数量分为 1-5，6-15，15 个词以上进行实验，实验结果如表 3 所示。

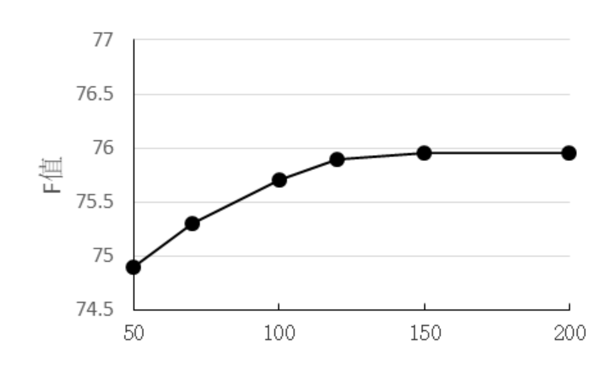


图 4　不同隐层大小句法分析结果

Fig. 4　Parsing results of different hidden layer sizes

表 3　不同句子长度句法分析结果

Table 3　Parsing results for different sentence lengths

句子长度	P/%	R/%	F/%
1-5	79.7	78.5	79.1
6-15	77.6	75.8	76.7
>15	73.4	71.8	72.6

根据实验结果可以看出句子长度对句法分析性能有较大影响，短句子的句法分析效果相较于长句子的效果较好，主要原因是短句子中短语的嵌套较少，句法树结构较为简单，因此句法分析效果较好。

4.3.3 句法分析器性能比较

为了比较本文提出的句法分析模型性能，与目前已有的哈萨克语句法分析模型进行了对比。本文提出句法分析模型的神经网络结构相关参数设置如表 4 所示。

表 4　神经网络结构参数

Table 4　Neural network structure parameters

神经网络结构	
词向量维度	50
词性向量维度	20
LSTM 层数	2
LSTM 隐层维度（每个方向）	120
ReLU 隐藏层单元个数（每个动作）	120

训练参数设置如表 5 所示。

表 5　训练参数设置

Table 5　Hyperparameters

训练参数设置	
向量初始化	随机初始化
迭代次数	10
梯度下降速率	p=0.5
自适应学习率	$\rho = 0.99$　$\varepsilon = 1 \times 10^{-7}$

按照表 4、5 的参数设置，将训练出的句法分析模型进行测试，所得的结果与文献[7~9]提及的句法分析结果进行了对比如表 6 所示，

表 6　不同句法分析器结果比较

Table 6　Comparison of performance of different parser on Kazakh test set

模型	P/%	R/%	F/%
文献[7]方法	68.34	67.47	67.90
文献[8]方法	70.27	69.39	69.83
文献[9]方法	71.4	69.6	70.5
动态规划	76.9	74.8	75.8

通过表 6 不同方法实验结果分析，基于句子跨度的哈萨克语短语句法分析与动态规划结合对哈萨克语短语句法分析结果有了进一步提升，相比与 PCFG 方法、PChart、PCFG 与感知机结合准确率分别提高了 8.56%、6.63%、5.5%。使用双向 LSTM 对句子跨度特征的提取在哈萨克语短语句法分析中起到了良好的效果。

5　结束语

本文研究了基于句子跨度的哈萨克语短语句法分析，相比于以往使用固定的特征模板获取句子上下文信息特征，使用了双向 LSTM 来获取句子跨度的上下文信息特征，在句法分析的准确率上有了一定的提高，并且将动态规划技术加入了哈萨克语短语句法分析上，取得了良好的效果。在将来的哈萨克语短语句法分析上，进一步增加训练语料的数量，采用不同的神经网络进行对比研究，进一步提高哈萨克语短语句法分析的准确率。

参考文献：

[1] Allen J F. Natural Language Understanding [M]//Natural Language Understanding. [S.l.]:Benjamin Cummings Pub. Co.,1995: 617-626.

[2] Colins M. Three generative, lexicalised models for statistical parsing [C]//Proc of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1997: 16-23.

[3] Hall D, Durrett G, Dan K. Less Grammar, More Features [C]//Proc of Meeting of the Association for Computational Linguistics.Stroudsburg, PA:Association for Computational Linguistics, 2014: 228-237.

[4] Charniak E. Statistical parsing with a context-free grammar and word statistics [C]//Proc of the 14th National Conference on Artificial Intelligence and the 9th Conference on Innovative Applications of Artificial Intelligence.AAAI Press,1997: 598-603.

[5] Dyer C, Kuncoro A, Ballesteros M, *et al.* Recurrent neural network grammars[C]//Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 199-209.

[6] Stern M, Andreas J, Dan K. A minimal span-based neural constituency parser [C]//Proc of Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 818-827.

[7] 尚文清，古丽拉•阿东别克，牛娜等. 基于 PCFG 模型的哈萨克语句法分析 [J]. 现代计算机，2015(5): 7-10. (Shang Wenqing, Gulila ALTENBEK, Niu Na *et al.* Syntactic Analysis of Kazakh Language Based on PCFG Model [J]. Modern Computer, 2015(5): 7-10.)

[8] 尚文清，古丽拉•阿东别克，牛娜等. 基于 PChart 算法的哈萨克语句法分析 [J]. 计算机工程与设计,2016,37(3): 832-836. (Shang Wenqing, Gulila Altenbek, Niu Na *et al.* Kazakh syntactic analysis based on probabilistic Chart algorithm [J]. Computer Engineering and Design, 2016, 37(3): 832-836.)

[9] 梁金莲，古丽拉•阿东别克. 由粗到精的哈萨克语短语结构句法分析研究 [J]. 中文信息学报，2018,32(1): 83-88. (Liang Jinlian, Gulila Altenbek. A coarse-to-fine Kazakh PSG parse [J]. Journal of Chinese Information Processing, 2018, 32(1): 83-88.)

[10] Hall D, Durrett G, Dan K. Less grammar, more features [C]//Proc of Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2003: 228-237.

[11] Cross J, Huang L. Span-based constituency parsing with a

structure-label system and provably optimal dynamic oracles[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2016: 1-11.

[12] Chen D, Manning C D. A fast and accurate dependency parser using neural networks [C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2014: 740-750.

[13] Cross J, Huang Liang. Incremental parsing with minimal features using bi-directional LSTM [EB/OL]. (2016-08-07). https://www.aclweb.org/anthology/P/P16/P16-2006.pdf.

[14] Mi Haitao, Huang Liang. Shift-reduce constituency parsing with dynamic programming and POS tag lattice [C]//Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1030-1035.

[15] Goldberg Y, Nivre J. Training deterministic parsers with non-deterministic oracles [EB/OL]. (2013-10-12). http://aclweb.org/anthology/Q/Q13/Q13-1033.pdf.

[16] Ballesteros M, Goldberg Y, Dyer C, *et al*. Training with exploration improves a greedy stack-LSTM parser [C]. (2016-10-27). http://www.aclweb.org/anthology/D/D16/D16-1211.pdf.

[17] Neubig G, Dyer C, Goldberg Y, *et al*. DyNet: the dynamic neural network toolkit [EB/OL]. (2017-01-15). https://arxiv.org/pdf/1701.03980.pdf.